# Self-Supervised Speech Representation Learning: A Review

Year: 2022 | Citations: 432 | Authors: Abdel-rahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob Drachmann Havtorn, Joaki

---

## Abstract

Although supervised deep learning has revolutionized speech and audio processing, it has necessitated the building of specialist models for individual tasks and application scenarios. It is likewise difficult to apply this to dialects and languages for which only limited labeled data is available. Self-supervised representation learning methods promise a single universal model that would benefit a wide variety of tasks and domains. Such methods have shown success in natural language processing and computer vision domains, achieving new levels of performance while reducing the number of labels required for many downstream scenarios. Speech representation learning is experiencing similar progress in three main categories: generative, contrastive, and predictive methods. Other approaches rely on multi-modal data for pre-training, mixing text or visual data streams with speech. Although self-supervised speech representation is still a nascent research area, it is closely related to acoustic word embedding and learning with zero lexical resources, both of which have seen active research for many years. This review presents approaches for self-supervised speech representation learning and their connection to other research areas. Since many current methods focus solely on automatic speech recognition as a downstream task, we review recent efforts on benchmarking learned representations to extend the application beyond speech recognition.